Psychological Monographs

General and Applied

Evaluating the Educability of the Severely Mentally Retarded Child

By

Helen Schucman

Division of Clinical Psychology,
Columbia-Presbyterian Medical Center

Price $1.00

Edited by Norman L. Munn
Published by the American Psychological Association, Inc.

Psychological Monographs:

General and Applied

Combining the *Applied Psychology Monographs* and the *Archives of Psychology*
with the *Psychological Monographs*

NORMAN L. MUNN, Editor

Department of Psychology, Bowdoin College
Brunswick, Maine

Manuscripts and correspondence on editorial matters should be sent to the Editor. *Psychological Monographs* publishes comprehensive experimental investigations and other psychological studies which do not lend themselves to adequate presentation as journal articles. Major space is given to the author's original contribution; introductory and bibliographic materials, as well as statistical tables and graphs, must be kept within reasonable bounds. Tables, graphs, and appendix materials which deal with detail not essential to adequate presentation of the findings may be made available through the American Documentation Institute—for details of this procedure, see the APA *Publication Manual.* Preparation of manuscripts for publication as monographs should follow the procedure given in the APA *Publication Manual.* Publication in *Psychological Monographs* is free of cost to the author, except in cases where early publication is requested or author's alterations are made in galley proofs.

ARTHUR C. HOFFMAN
Managing Editor

HELEN ORR
Promotion Manager

Correspondence on business matters should be addressed to the American Psychological Association, Inc., 1333 Sixteenth St., N.W., Washington 6, D.C. Address changes must arrive by the 10th of the month to take effect the following month. Undelivered copies resulting from address changes will not be replaced; subscribers should notify the post office that they will guarantee third-class forwarding postage.

Vol. 74, No. 14 Whole No. 501, 1960

Psychological Monographs: General and Applied

EVALUATING THE EDUCABILITY OF THE SEVERELY MENTALLY RETARDED CHILD[1]

HELEN SCHUCMAN[2]

Division of Clinical Psychology, Columbia-Presbyterian Medical Center

THE purpose of this study was to investigate a method for obtaining an objective, quantified, predictive measure of educability for children who are severely retarded mentally. The basic hypothesis was that the child's educability can be inferred from his responses to learning situations which require abilities on which education depends, namely, to learn from instruction, to transfer the training, and to retain the learning.

With increasing attention focused on the education of these children, the need for suitable methods of evaluating their abilities has become acute. Their limited abilities must be assessed, and attention directed toward developing the most profitable ones, to whatever extent this is possible.

Unfortunately, it is precisely these children whose abilities are among the most difficult to evaluate by the usual testing procedures. On standard tests, the child may be heavily penalized by verbal, motor, sensory, and experiential handicaps. Further, inability to cope with standard tests is one of the outstanding behavioral characteristics of the mentally retarded. Tests with which a child cannot cope will hardly serve to identify his assets. No curriculum can be built on what a child can*not* do.

It has been said that "the ideal test in the study of mental deficiency would be one which investigated the ability to learn" (Penrose, 1934, p. 49). This research was undertaken to study learning in severely retarded children, and to use the results as the basis for determining what they *can* do.

METHOD

The general method of testing was as follows. A battery of tests was administered to the subjects individually. Each subject was first given the tests without training. He was then trained to the correct responses according to standard training procedures, and retested to permit measurement of posttraining learning gains. Different forms of the tests were used to evaluate transfer. Retention scores were obtained by retesting the subject at various later dates.

Preliminary Experimentation

The final test battery was determined on the basis of the results of a series of pilot studies. The chief aim of the latter was to bring the tests

[1] The research reported herein was performed by New York University, School of Education, pursuant to Contract SAE-7783, with the Office of Education, United States Department of Health, Education, and Welfare.

Grateful acknowledgement is made to the following faculty members of New York University: Alonzo Grace, Howard Newburger, Roscoe Brown, Kenneth Benson, Jacob Cohen, and members of the Department of Educational Psychology; to Harold Michal-Smith of the New York Medical College, Christopher De Prospo of the College of the City of New York, Sidney Weinstein of Van Etten and Jacobi Hospitals, Marion Reissenweber of the Institute for the Crippled and Disabled in New York, and William Thetford, Arthur Carr, and Norman Reiss of the Columbia-Presbyterian Medical Center; and to the Mental Retardation Clinic at Flower and Fifth Avenue Hospitals in New York, the National Association for Retarded Children, the New York Association for the Help of Retarded Children, Parents and Friends of Retarded Children of Bridgeport, Connecticut, the Shield of David Institute for Retarded Children in New York, and the Westchester School for Retarded Children.

[2] Formerly at New York University.

into increasing accord with criteria which were established in advance of the preliminary experimentation. The criteria were as follows:

1. Tests should be appropriate to the level of ability of the subjects.

2. Tests should arouse and maintain the interest of the subjects.

3. Tests should be suitable for brain-injured subjects.

4. Tests should depent primarily on abilities sufficiently well developed at the age levels of the subjects.

5. Tests should depend minimally on past experience.

Before the pilot studies were undertaken, the following attempts to meet the criteria were made in the construction of the preliminary battery. To aid in meeting the first criterion, tests were designed so that no verbal responses were required, because of the verbal handicaps of the subjects. Instructions were given chiefly in pantomime, verbal directions being limited to simple words, and accompanied by explanatory gestures.

The literature was surveyed for areas suitable to the abilities of the subjects, and imitative abilities, memory, and gross shape, size, and color (brightness) discriminations were chosen. In addition to being adaptable to the subjects' abilities, these areas were thought to be relevant to the special purpose of the tests. Imitative teaching techniques and rote memory tend to be stressed in the training schools, because the children can grasp better what is shown or demonstrated than what is explained (Wallin, 1955, p. 391). On the other hand, simple size, shape, and color discriminations are often fundamental aspects of the curriculum. Further, since almost every child has had some opportunity to develop these discriminative abilities, tests which tap them seem to be among the most favorable measures of differences in ability (Stutsman, 1931, p. 45).

In the search for materials which would arouse the interest of the subjects, toys and training materials chosen by children in the training schools were noted. Sample materials were then constructed, and each sample was included in a group of toys presented to at least 12 severely mentally retarded children. Materials were considered only if they were spontaneously selected by more than half the children to whom they were offered. Short tests were planned, to hold the subjects' interest. The number of items in the individual tests ranged from four to eight, and were arranged in increasing order of difficulty.

Procedures used to meet the criterion of suitability for brain-injured subjects were in line with work done by Cotton (1939), Arnold (1945), Lord (1937), Strauss and Lehtinen (1950), Reissenweber (1953), Strauss and Kephart (1955), and Haeussermann (1958).

Materials were three-dimensional and relatively large. Surfaces were dull to minimize the reflection of light. Tests were not timed. The examiner manipulated the test materials at the subject's direction, if he could not do so himself. Only minimal responses, such as simple choices among the test materials, were required. Materials were presented far enough apart so that a gross motion of hand or arm sufficed to indicate the subject's selections. An achromatic scale was used, because of possible difficulties in color perception. Materials were painted black, white, and various grays in the Munsell Gray Scale, and were presented on a cloth of approximately the middle gray in the scale.

Developmental norms from the literature served as guides to aid in meeting the fourth criterion, and abilities usually developed at least one year below the age of the youngest subjects tested were selected. Two methods were used in the attempt to meet the requirement of minimal dependence on past experience. First, tests were sought which provided the experience necessary for performance in the testing and training procedures themselves. Otherwise, tests were considered if it was reasonable to assume that the materials and procedures were within the experience of all of the subjects.

Transfer situations were constructed in the light of E. L. Thorndike's theory, which holds that transfer depends on identical elements in the original learning and the new learning which it facilitates (Hilgard, 1948, p. 29). The transfer forms were developed in the course of the pilot studies by introducing various changes into the materials and/or the procedures used in the original forms of the tests. In order that the transfer forms would remain within the abilities of the subjects, only relatively few and comparatively simple changes were made. Some of the principles followed in the construction of the transfer forms were based on procedures used by Lashley (1938), Harlow (1949), Barnett and Cantor (1957), and House, Zeaman, Orlando, and Fischer (1957).

The sample used in the preliminary experimentation consisted of 115 mentally retarded children. Of these, 80 were between the ages of 5–0 and 12–6, and were severely retarded according to the criteria used in the selection of the final sample, as stated below. Seven pilot studies, each using 10 subjects, were conducted with this group. The remaining 10 were used for informal experimentation with various changes in materials and procedures, in advance of each pilot study.

In addition, eight severely retarded subjects between the ages of 3–6 and 4–1, and six children aged 12–7 to 14–0 were tested, to aid in determining the most suitable age range for the administration of the final battery. Ten moderately retarded subjects, and 11 whose retardation was attributed primarily to schizophrenia according to their medical and psychiatric records, were also tested on an experimental basis.

The original battery consisted of 11 tests. Preliminary work with some of these tests is reported by Schucman (1957-58). At least two tests in each area were included in the preliminary battery, since it was expected that some of them would prove to be unsatisfactory.

All of the tests were administered in the first four pilot studies. The less suitable ones began to be dropped thereafter. In the earlier pilot studies, only pretraining, posttraining, and transfer scores were obtained. The later ones also included retesting the subjects at various later dates, to aid in determining suitable time intervals at which to measure retention.

Several transfer forms for each test, based on changes in the original materials, in the procedures, or in both, were tried in the preliminary experimentation. The following procedures were used in introducing change into the materials. For tests of shape discrimination, brightness or size was altered, and shape was held constant. In size discrimination tests, shape or brightness was changed, and size kept constant. In test of brightness discrimination, brightness was held constant, and shape or size was changed.

Various alterations in the original procedures for administering the tests were also studied in the preliminary experimentation with the transfer forms. In one method, the original materials were presented, but in a different order. In a second method, the original demonstrations to the subject were included, but without the explanatory gestures which had accompanied them. In a third, the demonstrations were omitted, and only the explanatory gestures were given. In a fourth, the test was administered according to all of the original procedures, but in a different place.

Materials, procedures, and the order in which the tests were to be administered were determined in advance of each pilot study, and a list of criticisms was drawn up afterwards. Tests were changed, substituted, accepted, or rejected on the basis of their adequacy in meeting the test criteria. Quantitative and qualitative analyses of the data were made after each pilot study, to indicate strengths and weaknesses in the tests. The preliminary battery, and its evolution in the course of the pilot studies, are described in Appendix B.

Selection of Subjects for the Final Battery

The sample to which the final test battery was administered was limited to subjects suffering from various pathological physical conditions known to produce mental retardation, as stated in their medical records. Most of these records were obtained from special clinics for the mentally retarded. Some were supplied by physicians on the staff of the school which the child attended. In a few cases, the report of the child's personal physician was used.

Medical diagnoses ranged from relatively clear-cut statements of diagnostic category (such as "mongoloid") to more general, descriptive statements (such as "brain-injured"). Subjects whose retardation was ascribed primarily to "emotional difficulties" or schizophrenia, and categories such as "familial," "undifferentiated," "unclassified," and "etiology undetermined" were excluded from the sample.

In limiting the sample to severely retarded subjects on the basis of psychological test data, only evaluations made within the year, by qualified psychologists attached to the clinic or school which the child attended, were used. An IQ below 50 on the Revised Stanford-Binet Intelligence Scale was used as the criterion for all of the subjects to whom the scale could be meaningfully administered. Sixty children fell into this category. For subjects to whom the Binet scale could not be given, an IQ below 50 on any other standard intelligence test, such as the Merrill-Palmer Scale, the Cattell Intelligence Scale for Infants and Young Children, and the Kuhlmann-Binet Scale, was accepted. This criterion was used for an additional 41 subjects.

An estimated IQ below 50 was used for the 13 subjects who could not, or did not, respond to any of the standard intelligence tests in a way that permitted meaningful scoring. An estimated IQ was a measure made by the staff psychologist of the school or clinic, who had evaluated the child at least once during the past year, and who was personally familiar with him. The psychologist based the estimate on a study of the child's psychological record, which included the results of whatever standard test could be administered to the particular child, analyses and interpretations of the scores obtained, observations of the child's behavior, and clinical impressions of his level of mental functioning. The Vineland Social Maturity Scale, administered to the parent, was the only standard test that could be used for all of the subjects.

The age range for the final sample was set at 5–1 to 11–11. Results obtained in the pilot studies suggested that above that age level the tests tended to become too easy to permit discrimination among the subjects, and below it, the tests tended to be too difficult.

Screening Procedures for Sight and Hearing. The subjects' vision and hearing were evaluated by the following procedures. Medical records, teachers' reports, psychological evaluations, and questionnaires filled in by the parents were studied. Trained personnel familiar with the child were consulted. If these sources gave no indications that the child's sight and hearing were too seriously impaired, the examiner undertook special screening procedures for the specific purposes of the tests.

Before the child was brought into the room, a number of the smaller test materials were placed on the table, covering approximately the area to be used in the testing. If the child picked up these materials, reached toward them, or glanced from one to another, it was assumed that his sight was adequate for test purposes. The child's responses to remarks made by the examiner from various positions behind the child were also noted. Materials involving sound were introduced at several points outside his range of vision. If he turned, reached, or glanced in these directions, it was assumed that his hearing was adequate for the purposes of the tests.

Sample. The sample consisted of 114 mentally retarded children drawn from the Greater New York area, living in the community, and attending special classes and/or clinics for the mentally retarded. Their estimated IQs[3] ranged from 8 to 55, with a mean of 36.18 and a standard deviation of 12.01. They had attended school from 0 to 36 months, mean 11.52, standard deviation 10.84. Their ages ranged from 61 to 143 months, mean 95.16, standard deviation 23.48. Thirty-seven were mongoloid, and 11 were cerebral-palsied. The remainder suffered from various pathological conditions known to produce mental retardation, such as microcephaly, hydrocephaly, cretinism, and congenital cerebral hypogenesis. There were 69 boys and 45 girls.

Final Battery

The five tests which comprised the final battery are briefly described below. More detailed descriptions of materials and testing and training procedures are given in Appendix A.

I. Two-Buttons Test was primarily a test of imitative ability and memory. Materials consisted of a number of buttons, and two transparent boxes into which the buttons were dropped in alternating order. Any gross motion in the direction of the correct box or boxes constituted a correct response on the part of the subject. The test included six items, the buttons being presented to the subject in increasing numbers. Before each item, the examiner demonstrated the correct response. The subject was then requested to imitate the demonstration. In the transfer situation, the demonstrations were omitted, and only the original explanatory gestures were made.

This test was found to be a good beginning for the battery. A few preliminary clicks of the buttons in the boxes usually aroused immediate interest. Giving the child the buttons to play with, in advance of the test, was an excellent way of establishing rapport.

II. Buzz Test was chiefly a test of size discrimination. It consisted of a number of round, black discs of various sizes, attached to a gray board. Separate black discs of comparable sizes were presented to the subject, who was shown, in pretest demonstrations, that placing a disc over the corresponding sample on the board produced a continuous, buzzing sound. The subject was required to indicate the correct placement on the board for each of the discs presented to him. There were six items, arranged in increasing order of difficulty. In the transfer form, white discs, corresponding in sizes to the black ones, were presented to the subject, to be matched to the black samples on the board.

This proved to be a satisfactory second test in the battery. The "buzz" was sufficiently different from the sounds in the preceding test to evoke fresh interest and curiosity. Also, as one of the shorter tests, it was well suited to a period in which most of the subjects were still "settling down."

III. Music Box Test was largely a test of brightness discrimination. Round music boxes of equal size were used, one white, one black, one gray, one black and gray, and one gray and white. The subject was shown, in pretest demonstrations, that only the gray box "made music." Selection of the gray music box, made in any unambiguous way at the disposal of the subject, was the only response required. There were six items. The gray box was placed in a different position for each item, and the other boxes were added in increasing numbers. Larger, rectangular boxes, comparable in other respects to the round ones, were used for the transfer form.

This was the longest test in the battery. It was found to be most satisfactory when given in middle place, at a time when the subjects were usually sufficiently "settled," and when restlessness tended to be low.

[3] Estimated IQs up to 55 were used, provided the subject's score on the Stanford-Binet scale was below 50. Estimated IQs tended to be several points higher than Stanford-Binet IQs.

IV. Light Test, primarily a test of shape discrimination, consisted of a gray box with a plastic-covered window in each side. Knobs of various shapes appeared on each of the sides, arranged in increasing numbers. The subject was shown that only the rectangular knobs, one of which appeared on each side of the box, lit up a series of flashlight bulbs placed behind the windows. A correct response required the selection of this knob, made by the subject in any way possible to him. The test included eight items, each of the sides of the box being presented twice. A second gray box, with black knobs, corresponding in shape to the original white ones, constituted the transfer form.

As one of the shorter tests, this was chosen to follow the longest one in the battery. It was perhaps the best-liked of the tests, and was well suited to overcome developing restlessness and to restore flagging interest toward the end of the battery.

V. Three-Buttons Test, a second test of brightness discrimination, involved sorting white, gray, and black buttons. The examiner dropped one sample button into each of the three transparent containers. The subject was then requested to indicate the box which contained the sample corresponding to each of the buttons presented to him. The test contained six items, the buttons being presented in increasing numbers. In the transfer form, the positions of the sample buttons were reversed, and the subject was required to place the buttons according to the new pattern.

This test was placed at the end of the battery largely because a marked tendency toward perseveration resulted unless the two "buttons" tests were widely separated. Of the two, this one required less time to administer, and made a short, interesting final test.

Test Administrations

The final battery was administered nine times. Administrations A and B measured pre- and posttraining performance, respectively. The transfer forms were used in Administration C. These three administrations were completed for each test before the next one was begun. The subject was returned to his teacher or parent for the interval of approximately an hour which elapsed before the next administration was given.

Administration D measured retention without additional reinforcement. Pretest demonstrations and instructions were omitted, training was not repeated, and transfer forms were not used. Intervals of more than an hour for this method of retesting were found in the pilot studies to produce scores which were too low for purposes of discrimination. All of the subjects received the first four administrations of the battery.

Administration E was given approximately one week after the original testing. Training periods and the use of the transfer forms were again omitted. However, pretest demonstrations and instructions were included, since it was found that after an interval of a week the majority of the subjects in the pilot studies no longer remembered what was required without this amount of reinforcement. Of the original sample, 102 subjects were available for Administration E.

Administrations F, G, H, and I were given approximately seven weeks after the first testing period. They were repetitions of Administrations A, B, C, and D, respectively. Training was again included, and both forms of the tests were used. Thirty-six of the original subjects were available for the final four administrations of the battery.

Administrations A, B, and C of the battery, and the comparable administrations in the later retesting, required an average of about 40 minutes to complete. Administrations D and I could usually be given in a quarter of an hour. An average of about 20 minutes was needed for Administration E.

General Testing Procedures

Testing in the pilot studies was done by the investigator. Qualified, advanced graduate student testers administered the final battery. They were given a manual, and were supervised by the investigator in administering the tests with at least 10 subjects. These scores were not included in the results.

Subjects were tested individually, in the school or clinic which they attended. The number of sessions used to complete the battery depended on the needs of the particular child. Rest periods were sometimes necessary during the first three administrations. They were allowed only between the completion of Administrations A, B, and C of a particular test, and the beginning of the next one. If more than one session was necessary, the time intervals between the sessions were kept parallel when the child was retested. In most cases, however, the first three administrations were completed in a single session.

The child's handedness was determined in advance, if possible. In the relatively few cases where this was not possible, right handedness was assumed. Test items were presented toward the child's favored hand, the middle point of the item corresponding roughly to the midline of the shoulder on that side.

Since the examiner manipulated the materials for the child, if necessary, the subjects were given equal opportunities to see the completed responses, and to make or indicate spontaneous corrections. All responses, both in testing and training, were corrected or correctly completed by the examiner, so that the subjects had equal opportunities to see the correct response the same number of times.

General Training Procedures

It was found in the pilot studies that training the children to both the correct responses and to the avoidance of incorrect ones tended to be too confusing. Therefore, for the final battery, the subject was trained to the correct responses only. In order that the majority of the subjects would succeed in the training periods, extremely simple procedures were followed.

In general, training consisted of demonstrations of the correct responses by the examiner; imitation of this demonstration by the subject, with cues supplied by the examiner; and prevention or immediate correction of errors. Most of the subjects were able to achieve a number of correct responses under these conditions. The items were presented for training in the same order in which they were given in the tests. The subject was trained to a maximum of three times on each item, beginning with the first item failed on the test. Exceptions to these procedures are noted in Appendix A.

Scoring Procedures

The subject was given one point for each correct response. Scoring booklets contained one page for each test, on which each administration was scored separately. To minimize possible contamination in scoring, the original booklets were taken from the testers after the completion of Administration E. Also, the testers were not originally informed that any further testing was anticipated at that point. New scoring booklets were used for Administrations F, G, H, and I.

Teachers' Ratings

Ninety-nine of the subjects attended special schools for mentally retarded children. Teachers' ratings were obtained for this group, from the child's current teacher in whose class he had been for a period of at least one month. A five-point rating scale was used, on which the teacher was asked to rate the child's learning ability, as compared with that of the other children who attended that school. The various populations of the different training schools were thought to be sufficiently similar to warrant the use of this frame of reference.

Results[4]

Test-Retest Comparisons

In order to determine the stability of the battery, product-moment correlations were computed between the scores obtained on comparable administrations of each test, by the 36 subjects who were still available for retesting after an interval of seven weeks. The coefficients, all of which were highly significant, are shown in Table 1.

TABLE 1

Product-Moment Correlation Coefficients* Obtained by Test-Retest Comparisons
($N = 36$)

Administrations	Test I	Test II	Test III	Test IV	Test V
A and F	.66	.82	.85	.79	.80
B and G	.93	.95	.94	.92	.95
C and H	.93	.90	.91	.92	.96
D and I	.95	.90	.95	.92	.95

* All significant at .001 level.

It can be seen that the coefficients produced by the pretraining administrations of the tests (Administrations A and F) were always lower than those obtained between comparable learning scores. The former ranged from .66 to .85 and the latter from .90 to .96, with no overlap between the ranges. The likelihood that the A-F coefficients would be consistently lower than those between the learning scores is exceedingly small ($p < .01$) by the Mann-Whitney U test (Siegel, 1956).

[4] Statistical tables, referring to the original investigation and entitled ***Supplement to Measurement of the Educability of Severely Mentally Retarded Children*** (Schucman, 1959), are available on interlibrary loan from: Gould Memorial Library, New York University; New York 53, New York.

Intertest and Intratest Comparisons

To determine the intertest consistency of the battery, product-moment correlations were performed between the scores from the same administrations of the different tests, i.e., scores from Administration A of each test were compared with scores from Administration A of every other test, and so on. The 102 subjects who remained available for Administration E were used in these analyses, which were based on scores from Administrations A, B, C, D, and E of each test. The correlations thus resulted in 50 coefficients, 10 for each of the administrations.

All of the coefficients reached the .001 level of confidence. However, those yielded by the pretraining scores, which ranged from .31 to .63, were significantly lower than those produced by the 40 comparisons between the learning scores, which ranged from .72 to .87, with no overlap between the ranges ($p < .01$ by the Mann-Whitney U test).

To determine the intratest consistency of the battery, product-moment correlations were also computed between the scores obtained from different administrations within the same tests, i.e., between scores from Administrations A, B, C, D, and E of each test, using 102 subjects. Fifty coefficients, 10 for each of the tests, were thus obtained.

Again, all of the coefficients were significant at the .001 confidence level. However, the 20 coefficients obtained between the pretraining and the various learning scores within each of the tests were consistently lower than those produced by the 30 correlations of the various measures of learning performance. The former ranged from .50 to .89 and the latter from .86 to .97, with overlap between the ranges occurring in only one instance ($p < .01$ by the Mann-Whitney U test).

Comparisons of Scores Obtained on Different Administrations

In order to determine whether the different administrations of the battery had produced significantly different results, a series of subjects × conditions analyses of variance (Edwards, 1950) were performed, comparing pretraining, posttraining, transfer, and retention performance of all of the subjects, as obtained on Administrations A, B, C, and D of each test.

The results demonstrated that highly significant differences were obtained on every test, the *F*s ranging from 94.90 to 163.97 ($df = 3, 339$). With a single exception (between posttraining and retention scores on Test III), the *t*s between the means, which ranged from the .05 to the .001 levels of confidence, showed that significant differences were found between the scores from all of the administrations of all of the tests. The mean scores obtained on the various administrations of each of the tests are shown in Figure 1. (In comparing the mean scores from the different tests, it should be borne in mind that Test IV contained eight items, while the others included six).

A second series of analyses of variance was performed between the scores obtained on comparable administrations by the 36 subjects who received all of the administrations of the battery. The *F*s between administrations ranged from 12.16 to 29.44 ($df = 7, 245$), all significant at the .001 level of confidence. The *t*s between the

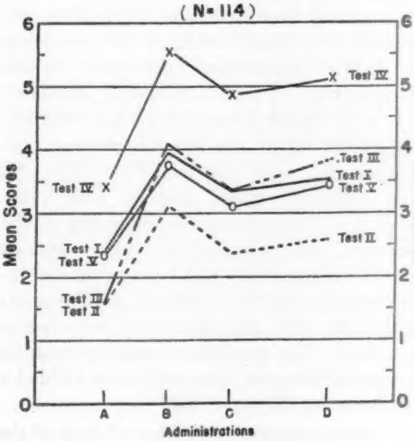


FIG. 1. Differences in mean scores obtained on various administrations.

means obtained on the different administrations demonstrated the following.

Except on Test 1, the scores produced by Administration F were significantly higher than those from Administration A, with *t*s from the .05 to the .001 levels of confidence. Administration F always produced significantly lower scores than Administration B, with highly significant *t*s. In three out of five instances, scores from Administration F were also significantly lower than those from Administration D. It was also found that similar learning curves were obtained across time by Administrations A, B, C, and D on the one hand, and Administrations F, G, H, and I on the other.

Effect of Population Variables on Test Performance

The influence on test performance of five independent variables, namely, sex, time spent in school, age, diagnostic category, and intelligence, was analyzed. Estimated IQs were used as the measure of the subjects' intelligence, since no one standard intelligence test could be validly administered to all of the subjects. Estimated IQs were thought to be the most suitable measure of intelligence for the group for the following reasons.

Estimated IQs, Vineland SQs, teachers' ratings, and Stanford-Binet IQs where obtainable were intercorrelated (r), resulting in highly significant coefficients which ranged from .54 to .92. However, estimated IQs were found to have the closest relationships with all of the other measures, correlating .92 with Stanford-Binet IQs, .82 with teachers' ratings, and .81 with Vineland SQs. To estimate the reliability of the measure, two estimated IQs were obtained for 11 subjects who had been evaluated independently by two different psychologists within the year, once at a clinic and once at a school. The product-moment correlation computed between these estimates yielded a coefficient of .87 ($p < .001$).

Before analyzing the effect of each of the five population variables on test performance, multiple correlations (R) were performed, to determine the extent to which these factors together accounted for the variation in the scores. All of the subjects, and scores from Administrations A, B, C, and D of each test, were used in these analyses.

The coefficients, all highly significant, are shown in Table 2. It can be seen that the coefficients between the pretraining scores and the independent variables, which ranged from .48 to .78, were always lower than those between the learning scores and the independent variables, which ranged from .83 to .90, with no overlap between the ranges ($p < .01$ by the Mann-Whitney U test).

TABLE 2

MULTIPLE CORRELATION COEFFICIENTS* BETWEEN THE FIVE INDEPENDENT VARIABLES AND ADMINISTRATIONS OF EACH TEST ($N = 114$)

Administration	Test I	Test II	Test III	Test IV	Test V
A	.72	.72	.48	.67	.78
B	.88	.87	.86	.86	.87
C	.90	.85	.86	.85	.85
D	.83	.84	.84	.86	.86

* All significant at .001 level.

Expressed as coefficients of determination (R^2), the above results demonstrate that the factors of sex, time in school, age, diagnostic category, and intelligence together accounted for 23 to 61% of the variation in pretraining performance, and 69 to 80% of that obtained in the learning scores.

The influence of each of the five population variables on the scores obtained from the first five administrations of each test was analyzed by a series of "repeated measurements" analyses of variance (Edwards, 1950). Groups were proportionalized so that nonsignificant differences existed between them with regard to each of the four population variables not under consideration at the time. Chi square and *t* tests were used to test the significance of these differences.

Effect of Sex. Analyses for determining the influence of the factor of sex on the test scores were based on 24 boys and 24 girls. The results demonstrated that this variable did not significantly affect performance on any of the tests, either as a main effect or in interaction with the administrations.

Effect of Time in School. The three levels of time spent in school which were used in the analyses for the effect of this variable on the test scores were based on the length of time of the subjects' school attendance. Nine months were counted for each actual year, the vacation months being excluded.

Subjects with little or no schooling were assigned to the lowest level. There were 28 subjects in this group, who had attended school from 0 to 2 months, with a mean of 1.04. The middle level consisted of 26 subjects, whose school attendance ranged from 3 to 13 months, mean 7.58 months. The 21 subjects in the highest level had attended school for relatively long periods, ranging from 14 to 33 months, mean 21.53.

The results of the analyses demonstrated that the factor of time spent in school did not significantly influence performance on any of the tests, either alone or in interaction with the administrations.

Effect of Age. The levels used for the variable of age corresponded to three general levels of school readiness for this type of population. The groups were somewhat older than is customary with "normal" children at corresponding levels, which is usually the case with severely mentally retarded subjects.

The youngest level corresponded to a preschool group, and consisted of 37 subjects, aged 61 to 83 months, with a mean of 71.00 months. The middle level constituted a beginning and early school group, and contained 36 subjects whose ages ranged from 84 to 112 months, mean 96.39. The 21 subjects in the oldest level were ready, on the basis of age, for more sustained and concentrated training. Their ages ranged from 115 to 143 months, mean 126.81.

Analysis of variance demonstrated that age did not significantly affect performance on Tests I, IV, or V, and no significant interaction effect was found on any of the tests. However, the *F*s between the age levels on Tests II and III were both significant at the .05 level of confidence, showing that the factor of age exerted a significant effect on the scores of each of these tests considered as a whole. The between-level *t*s ($p < .05$) demonstrated that the mean score of the youngest group was significantly lower than that of the middle or oldest levels on both of these tests. No significant differences were obtained between the scores of the middle and oldest groups, or between the age levels on specific administrations.

Effect of Diagnostic Category. Levels used for the variable of diagnostic category were based on the extent of the child's handicaps and the impairment of his general level of functioning. The 55 brain-injured subjects (i.e., children who had suffered trauma to the brain) were assigned to the lowest level. They were thought to be the most handicapped in the sense that various types of physical handicaps were present in addition to the mental retardation. Mongoloids were placed at the highest level, which included 17 subjects. They were considered to be the least handicapped in that, as a group, their motor coordination was relatively good, and their behavior comparatively stable. The middle level consisted of 21 subjects in various diagnostic categories occurring in insufficient numbers to be treated as separate groups, and whose medical records and life histories lacked evidence suggestive of trauma to the brain.

The analyses demonstrated that the factor of diagnostic category did not significantly influence the scores obtained on Tests I, II, III, or V. There was no significant interaction effect on any of the tests. However, an *F* significant at the .05 level of confidence on Test IV showed that this variable affected performance on the test as a whole. The between-level *t*s ($p < .05$) demonstrated that the mean score of the mongoloids was significantly higher than those of the other two groups. No significant differences between the levels of diagnostic category were obtained within specific administration.

Effect of Intelligence. Three levels were used for the factor of intelligence, based on the division of the total range of estimated IQs approximately into thirds. The lowest level contained 21 subjects, whose estimated IQs ranged from 8 to 25, mean 15.95. The middle level, which consisted of 31 subjects, ranged in estimated IQs from 27 to 39, mean 34.10. The estimated IQs of the 50 subjects assigned to the highest level ranged from 40 to 55, mean 46.40.

Analysis of variance clearly demonstrated that the factor of intelligence, both alone and in interaction with the administrations, exerted a highly significant influence on the scores of every test. These results are shown in Figure 2, which presents the mean scores obtained by the three levels of intelligence on the five administrations of each of the tests.

The *F*s obtained between the levels of intelligence, all of which were highly significant, ranged from 68.62 to 121.92 ($df = 2, 99$). The between-level *t*s, all significant at the .001 level of confidence, showed that the high intelligence group performed significantly better than the low or middle levels, while the low group was significantly inferior to the middle level on every test considered as a whole.

Further, the *t*s between the intelligence levels on specific administrations, significant at the .001 confidence level in every case, demonstrated that the mean scores of the high intelligence group were significantly greater than those of the low group in all types of learning performance. Also, the low group was always significantly inferior in learning performance to the middle group, and, with one exception, the middle intelligence level was significantly inferior to the high group.

It can therefore be said that significant differences were found between the performance of the three levels of intelligence in 59 of the 60 comparisons based on learning scores. On the other hand, the *t*s for pretraining performance, which were always lower than those obtained for the learning scores, failed to reach statistical significance in 4 of the 15 comparisons.

Of particular interest was the highly significant interaction between the levels of intelligence and the test administrations, which was found on all five tests. The interaction *F*s ranged from 7.01 to 20.95 ($df = 8, 396$). The within-level *t*s showed the following differences in the performance of the different levels on the various administrations of the tests.

The only significant differences found in the scores of the low intelligence level were obtained between pretraining and posttraining scores on four of the five tests. The group failed to transfer the training and

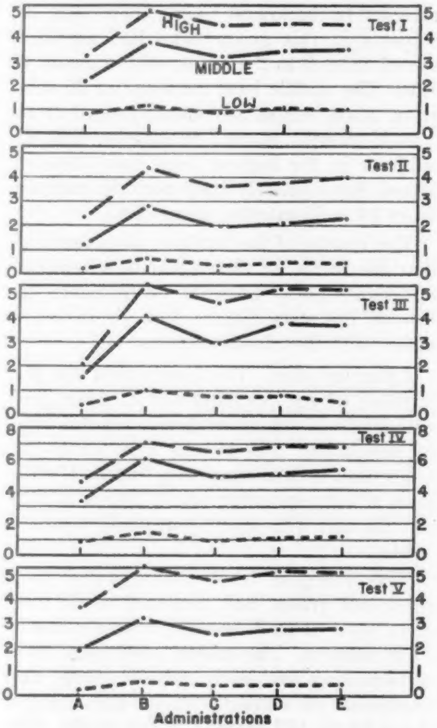


FIG. 2. Mean scores obtained at different levels of intelligence.

to retain the learning to a significant extent on any test.

For the middle intelligence level, all mean scores obtained for learning performance were significantly higher than those from pretraining administrations ($p < .001$). Also, significant differences were found in 17 of the 30 comparisons made between the various learning scores obtained by this group.

Highly significant differences were always found between pre- and posttraining scores for the high level of intelligence. The *t*s between the means were always larger than those for the middle intelligence level, with no overlap between the ranges ($p < .01$ by the Mann-Whitney *U* test). Among the different types of learning scores, highly significant differences were obtained in 21 of the 30 comparisons.

Still another demonstration of the significantly greater influence of the factor of intelligence on learning as opposed to pretraining performance was found in coefficients of determination based on multiple correlations (R^2). All of the subjects, and scores from Administrations A, B, C, and D of each test, were used in these analyses.

The results demonstrated that the variable of intelligence accounted for 18 to 55% of the variation in pretraining performance, and 62 to 76% of the variation obtained in the learning scores, with no overlap between the ranges ($p < .01$ by the Mann-Whitney *U* test).

Comparisons of Test Scores with External Criteria

The intercorrelations of estimated IQs, teachers' ratings, Vineland SQs, and Stanford-Binet IQs have already been mentioned in the section devoted to the relationships of estimated IQs with the other measures. The coefficients were all significant at the .001 level of confidence, demonstrating that the criteria were not unrelated. In addition to those already reported, the correlations produced a coefficient of .61 for Vineland SQs and Stanford-Binet IQs, and teachers' ratings correlated .72 with Vineland SQs and .54 with Stanford-Binet IQs, the latter being the lowest coefficient obtained.

To estimate the reliability of the teachers' ratings, a product-moment correlation was performed between independent ratings made by two teachers who had taught the same child for at least one month during the past year. Two such ratings were obtained for 25 of the subjects, and correlated .85 ($p < .001$).

Product-moment correlations were computed between the scores from Administrations A, B, C, D, and E of each test, and teachers' ratings, Vineland SQs, and psychologists' estimated IQs. Although the close relationship between estimated IQs and test performance was already demonstrated, this measure was again included here because of the added information supplied in this context.

The analyses resulted in 75 coefficients (5 administrations × 5 tests × 3 criteria). Those coefficients ranged from .31 ($p < .01$) to .87, all except the first reaching the .001 level of confidence. With two exceptions, both based on pretraining performance, the coefficients between the test scores and estimated IQs were always higher than those obtained between the test scores and the other criteria. In 18 out of 25 instances, the relationships between teachers' ratings and the test scores were greater than those between Vineland SQs and the test scores.

The correlations of pretraining scores and estimated IQs yielded coefficients ranging from .44 to .77, while those obtained between the learning scores and this criterion ranged from .70 to .87. Within each test, the coefficients based on learning scores were always higher than those obtained for pretraining performance.

To determine the significance of these differences, *t*s were found for the differences between the pairs of correlation coefficients, one of pretraining score with the criterion and the other of learning score with the criterion, as obtained for each test (Guilford, 1956, pp. 193f.). The results showed that the relationship between learning performance and estimated IQs was always significantly greater than that between pretraining performance and this criterion.

The correlations between teachers' ratings and the test scores produced the following results. The coefficients based on pretraining scores ranged from .47 to .72, and those obtained between the learning scores and this criterion ranged from .67 to .78. Again, the latter were always greater than the former within each test. The *t*s between these differences demonstrated that the relationships between the learning scores and teachers' ratings were significantly greater than those between initial test performance and this criterion in 16 of the 20 comparisons.

Similar results were obtained by the correlations of Vineland SQs and the test scores. The coefficients for initial test performance and the criterion ranged from .31 to .70, while those obtained for the learning scores ranged from .64 to .76, with no overlap within the separate tests. The *t*s performed between the coefficients showed that the learning scores were more significantly related to Vineland SQs than were the pretraining scores in 15 instances out of 20.

Discussion

Pilot Studies

Although the immediate purpose of the pilot studies was to increase the adequacy of the tests in meeting the criteria established for them, a number of ancillary findings were obtained from the preliminary experimentation which suggest further research in two main directions. The first is the possible extension of the method of testing used in this study to other groups of children. The following findings are relevant in this connection.

Several tests in the experimental battery, and some of the earlier forms of the final tests (see Appendix B), were found to be too difficult for the sample used in the present study, and were therefore not used in the collection of the final data. However, they were administered experimentally to 10 moderately retarded subjects aged 5–8 to 11–1, with IQs on the Revised Stanford-Binet Intelligence Scale ranging from 51 to 74.

Though the number of subjects was too small to warrant generalizations, the obtained results suggest that these tests are suitable for evaluating learning abilities in moderately retarded subjects, and may provide the basis for an upward extension of the present battery.

On the other hand, some of the tests in the original battery, as well as some of the early forms of the final tests, were found to be unsuitable for the present study because they were too easy for the subjects. These tests were tried with eight severely retarded children, aged 3–6 to 4–11. Although the number of subjects was again very small, the results indicate that these tests can be used for severely retarded children below the age levels used in this study, and may be a basis for a possible age scale.

Of considerable interest are the results obtained by 10 subjects diagnosed as schizophrenic on the basis of medical and psychiatric records. These subjects were among a group of 24 children, who had already been evaluated by a psychologist but whose medical records were not yet available. All of these children were given the final battery, and the results were set aside pending the arrival of the medical records.

Analyses of the scores showed that 11 of the subjects had performed markedly better on all of the tests than did the remainder of the group, whose mental levels were comparable. When the medical reports arrived, it was found that 10 of the 11 subjects were diagnosed as schizophrenic. It is possible that tests such as these may provide a method for differentiating the so-called "pseudoretardate" from the mentally retarded child functioning at a comparable mental level.

Results obtained on tests which were ultimately rejected as unsuitable for brain-injured subjects are also of special interest. Preliminary analysis of the limited results obtained in this area suggest that the brain-injured children were more sharply differentiated from the non-brain-injured subjects by their learning performance than by their initial test scores. It may be that learning behavior, as measured on tests such as these, is a better indicator of the presence of brain-injury than are single administrations of a test or tests.

The second area which warrants further investigation, on the basis of results of the pilot studies, is the possible application of the improvements made in the tests during the preliminary experimentation to curricular materials and training methods in the classroom.

Among the chief difficulties which arose in the development of the battery were the problems of increasing motivation for learning, improving the procedures for training, and minimizing the tendency toward perseveration. A number of successive changes were introduced into the test during the pilot studies, which progressively increased the efficiency of the tests in these respects. The specific set of variables responsible for these improvements cannot be identified at this time. However, the results are suggestive, and may have direct applicability to the training of severely mentally retarded children in the classroom. The findings are as follows:

1. Simple sensory rewards, such as the sound of bells, buzzers, and chimes, and the turning on of lights, were the most satisfactory devices for increasing motivation.

2. These rewards could be built into the materials themselves, or rewards should be supplied by the examiner verbally as the child performed. A combination of both was the most successful.

3. The positive effect of these rewards lasted for relatively short periods. Thereafter, the introduction of another such reward was necessary to reawaken interest and maintain it for another brief period.

4. The factor of novelty appeared to be more important in the rewards than was the particular sense modality involved, i.e., substituting one sound for another was as effective as changing from an auditory to a visual reward.

5. The use of external rewards, such as offers of toys or candy, was often helpful in establishing and maintaining rapport, but was not found to be a significant factor in motivating performance on the tests.

6. The usual testing practices (smiling, nodding, and so on) were highly unsatisfactory. An extremely active and constant participation on the part of the examiner, on the other hand, was a highly facilitating factor.

7. The participation of the examiner was most effective in the form of clapping, laughing, singing, and patting the child.

8. Training to the correct responses only was considerably more productive than training to both selection of the correct responses and avoidance of the incorrect ones. The inclusion of both tended to confuse the subjects.

9. When it was necessary to demonstrate to the subject that certain of the materials were nonfunctioning, it was essential to reinforce the correct response immediately thereafter.

10. The most suitable training methods were those in which all procedures were repeated without variation. Whether this would be so in longer training periods is not known. However, it can be said that the most effective combination for facilitating learning in the short training sessions used for the tests was that of constant changes in the materials, without the introduction of change into the training methods which were used with each of them.

11. Perseveration was minimized when the pattern of the required responses was not changed, and only the number of the responses was increased. The introduction of even minimal changes into the pattern tended to result in an increase in perseveration.

12. Items which required choosing among materials so arranged that the increment in size or degree of brightness between adjacent materials was constant tended to increase perseveration.

13. The tendency to perseverate was minimized when dissimilar objects were placed next to each other, and also when the test items were arranged so that the

selection of adjacent materials was not required.

14. Perseveration was apt to occur when training began with more complex items. Under these circumstances, the subjects were often unable to return successfully to the simpler responses which they had previously given correctly.

15. Single demonstrations of one or two of the simpler items, before undertaking training on the more complex ones, greatly facilitated a later return to the simpler items previously passed.

Adequacy of the Final Battery in Meeting the Test Criteria

The tests were apparently within the abilities of the subjects since, as a group, they obtained significantly higher scores after training, and were able to transfer the training and to retain the learning to a highly significant extent on every test. It can also be said that the tests were able to discriminate among the subjects, since children at different levels of ability produced significantly different scores.

The behavior of the subjects in the test situation clearly pointed to the adequacy of the battery in arousing and maintaining their interest. These qualitative observations were supported by the analyses of the quantitative results. The inter- and intratest comparisons demonstrated a high degree of consistency in the performance of the group, and a definite, consistent, and stable pattern was found in all of the learning curves. Such findings would hardly be likely to occur in the face of lack or loss of interest.

Although the effect of brain-injury on test performance requires further study, the obtained results suggest that the battery was essentially suitable for these children. The brain-injured group succeeded in making significant gains in immediate learning, and was able to transfer and retain the learning to a significant degree. Further, children with no speech and with severe motor handicaps were able to respond adequately.

The obtained results are also thought to show that the tests depended chiefly on abilities sufficiently well developed at the age levels of the group. Although a limited effect of the variable of age was found on two of the tests, significant learning gains were obtained by the different age levels on all of them.

Adequacy in meeting the requirement of minimal dependence on past experience can only be inferred. Perhaps the most directly relevant finding was that the factor of time in school, which ranged from none through five years, did not significantly influence performance on any of the tests.

Differential experiences are necessarily involved in a range of five years of school attendance. Further, school experiences are of a kind which might a priori be believed to be the most likely to influence performance on tests such as these, especially since the materials and procedures were not unlike those used in the training schools. It can therefore be said that precisely those kinds of experiences which might reasonably be expected to exert a differential influence on test performance failed to do so.

In brief, it is thought that the results demonstrated that the battery met the criteria reasonably well. The tests were appropriate to the range of abilities and the age levels studied, succeeded in arousing and maintaining the interest of the children, and were suitable for subjects with severe motor and verbal handicaps.

Effects of Population Variables on Test Performance

No significant influence of the factors of sex or time spent in school was found on the scores obtained from any of the tests. Apparently the battery was suitable for either boys or girls, and for children with school experience ranging from none to five years. The variable of age exerted a limited effect on the scores of some of the tests, its influence being restricted to the performance of the youngest subjects.

No significant influence of the factor of diagnostic category was obtained for the majority of the tests, and only a limited effect of this variable, restricted to the performance of the brain-injured subjects, was found at all. However, these results were to some degree a function of the levels into which the group was divided for the analyses in which the findings were obtained.

It is not thought that the levels of diagnostic category which were used in the present study were adequate for a thorough investigation of its possible effects. Although no rationale for a meaningful subdivision was found with this sample, the use of a brain-injured group as such was open to serious question, since brain-injury is a highly heterogeneous phenomenon. Additional study is also needed in connection with the various categories assigned to the miscellaneous group.

Intelligence was the only factor which was found to be highly related to performance on every test. Of particular interest in this connection are the results which demonstrated that a closer relationship was always obtained between intelligence and learning performance than between intelligence and initial test performance. The factor of intelligence not only exerted a highly differential influence on pretraining as opposed to learning performance, but also on the various types of learning scores obtained, i.e., immediate learning, transfer, and retention.

While subjects characterized by the lowest level of intelligence made significant gains in immediate posttraining learning on most of the tests, they did not succeed in transferring the training or in retaining the learning on any of them. Since the deficiencies of these subjects were most clearly demonstrated in the latter areas, it can be said that transfer and retention scores were the most sensitive of the learning measures to differences in ability.

In contrast to the performance of the subjects at the lowest level of intelligence, the middle intelligence group not only gained in response to training, but was also able to transfer the training and to retain the learning to a highly significant extent on every test. While the high intelligence group also made significant gains in the three learning areas, the magnitude of their mean scores was always higher than that obtained by subjects in the middle intelligence group.

Significant differences in the learning performance of subjects at different levels of intelligence were found in 59 out of 60 comparisons. On the other hand, the magnitude of the differences in their respective mean scores, as obtained by the pretraining administrations of the tests, was always lower than those found in their learning scores. Further, pretraining performance failed to differentiate the groups to a significant extent in 4 out of 15 instances.

These results carry the implication that for severely mentally retarded children, and perhaps for others as well, tests which are administered only once may not be the most sensitive measures of differences in ability.

Reliability

Test-retest comparisons demonstrated a high degree of stability for all of the tests. However, the learning scores on every test were found to be significantly more stable than the pretraining scores, with coefficients of .90 or higher in every case. The learning scores were also found to be significantly more closely related than were the pretraining scores to other learning trials on the same test, and to the same trials on different tests.

These findings suggest that initial test performance was responsive to more specific factors, possibly including previous experience. On the other hand, the greater stability and consistency of the learning scores suggest that a more general factor was measured.

Validity

The obtained results consistently attest to the construct validity of the tests. The evidence on its behalf began with the demonstrated adequacy of the battery in meeting the criteria established for the tests, which was a precondition of validity. In order to measure a child's abilities, he must be placed in situations which are appropriate to his age and to his level of ability, in which he is highly motivated to respond, and in which he is capable of responding within his limitations.

The basis for construct validity lay in the particular learning abilities which were studied. Educability can virtually be defined in terms of a child's ability to learn in response to instruction, to transfer the training, and to retain the learning. Schools are established in order to obtain the desired outcomes of learning, and "it might almost be said that if there is to be education there must be transfer" (Pressey & Robinson, 1944, p. 573).

The particular areas in which the learning abilities were investigated are also thought to contribute to the construct validity of the battery, in that imitative abilities are important in the training of these children, and abilities to make shape, color, and size discriminations are emphasized in the curriculum.

The shape of the learning curves, consistently obtained by the various tests and persisting across time, was also in close accord with the constructs. The lowest scores made by the group were always found in the pretraining administrations of the tests. On the other hand, the highest scores were always obtained from the administration which was given immediately after training, where one would assume that the training would have its maximum effect.

Also in accord with the constructs, the transfer scores made by the group were significantly higher than initial scores and lower than immediate learning gains, indicating that the group successfully transferred some of the training to other situations. Finally, the retention performance of the group was significantly higher than the original scores and lower than posttraining learning gains. Although the practice effect which the use of the transfer forms may have introduced is not known, and although the possible influence of each retest on subsequent ones was not determined, the learning was retained to a significant extent, without repetition of the actual training.

The validity of the constructs would also require that the measurements be stable over time, and, if more general learning abilities were tapped, that a high degree of consistency among the learning scores be present. Further, if learning behavior, as obtained on the tests, was related to actual differences in ability, the magnitude of the learning scores would be different at different levels of ability, and would increase as the level of ability increased. It would also be necessary that the learning scores be significantly more sensitive than the pretraining scores to these differences. These results were found on every test.

The results give evidence of predictive as well as of construct validity for the battery. The specific power of the tests to predict the ability of the child to learn in the classroom situation, which would be required if the battery were to be used for purposes of selection or for making actual educational decisions, remains to be demonstrated. Meanwhile, the obtained results offer indirect evidence that the tests would be useful for such purposes. This evidence was obtained in the comparisons of the test scores with the external criteria. Before these relationships are considered, the criteria themselves will be briefly discussed.

Teachers' ratings constituted the closest approach to a criterion of educability per se in this study, in that they reflected the actual experience of the teachers with the children's classroom learning over time. The ratings were made by teachers with special training and experience in the area of mental retardation, who were personally familiar with the children whom they rated. While subjective factors probably entered into their ratings, some amount of objectivity could reasonably be anticipated. The reliability of the ratings was indicated by the highly significant agreement found in the comparison between independent ratings of two teachers evaluating the learning performance of the same child.

A highly significant relationship was also found between the teachers' ratings and psychologists' estimated IQs, each being unaware of the other's estimate. While this agreement may show that the two measures supported each other, it must be borne in mind that the psychologists' records were available to the teachers, so that the extent to which the teachers were influenced by this information cannot be determined.

It can be said, however, that while making the actual ratings, the teachers reviewed their own longitudinal records of the child's performance in the classroom, and did not refer to the psychological records at that time. It can also be said that, although the ratings of the teachers were found to be significantly related to all of the other criteria, the agreement was not so close as to preclude the possibility of independent judgment.

Psychologists' estimated IQs, on the other hand, were found to be very closely related to the scores of standard tests, especially to IQs on the Revised Stanford-Binet Scale where obtainable. This close relationship, of course, may well be due to the fact that the psychologists based their estimates very largely on psychometric data in the children's records.

However, since some amount of subjectivity was probably unavoidable in the estimates of the psychologists, the highly significant relationships between this measure and the scores of standard tests were desirable. The reliability of the measure was indicated by the highly significant correlation between two such estimates made independently by different psychologists evaluating the same child.

The Vineland Social Maturity Scale is a measure of "social competence," which its author defines as "the fuctional ability of the human organism for exercising personal independence and social responsibility" (Doll, 1953, p. 12). The concept is directly relevant to the training of these children, since the curriculum places major emphasis on the development of traits which the construct reflects. This was the only standard test that could be used with all of the subjects, since a social age and a social quotient could be derived on the basis of data supplied by the parent, rather than by the subject himself.

The comparisons of the test scores with the external criteria demonstrated that the results obtained by the various administrations of the tests were significantly related to all of the criteria. However, a marked difference in the degree of the relationship was found for initial test performance on the one hand, and for the learning scores on the other.

In all cases, the correlations between learning performance and psychologists' estimated IQs were significantly greater than those obtained between estimated IQs and initial test scores. In most cases, the relationships found between learning performance and teachers' ratings were significantly greater than those between the pretraining scores and this criterion. Also in most cases, the learning scores were significantly more related to Vineland SQs than was initial performance on the tests.

It can therefore be said that the learning scores obtained were better predictors than were the initial test scores of intelligence, classroom learning, and social competence, as measured by estimated IQs, teachers' ratings, and Vineland SQs, respectively.

Summary and Conclusions

The purpose of the study was to investigate a method for obtaining an objective, quantified, predictive measure of educability for the severely mentally retarded child. The method was designed to measure abilities on which education depends, namely, to learn under training, to transfer the training, and to retain the learning.

The subject was first given a battery of tests without training. He was then trained to the correct responses under standard conditions, and retested so that learning gains could be measured. Different forms of the tests were used to study transfer, and retention scores were obtained by retesting the subjects at various later dates.

Criteria established for the adequacy of the tests required that they be appropriate to the abilities and ages of the subjects; capable of arousing and maintaining their interest; and suitable for their verbal, motor, perceptual, and experimental handicaps. Areas selected for study were imitative abilities, memory, and gross shape, brightness, and size discrimination.

A series of pilot studies was conducted with a preliminary battery which consisted of 11 tests. The chief aim of the pilot studies was to bring the tests increasingly into accord with the criteria. Tests were altered, substituted, accepted, or rejected on this basis.

The sample consisted of 114 severely mentally retarded subjects, living in the community, and attending special classes and/or clinics for mentally retarded children. They suffered from various pathological physical conditions known to produce mental retardation. Their estimated IQs (a measure made by the evaluating psychologist, based on quantitative and qualitative analyses of the child's psychological records) ranged from 8 to 55, mean 36.18, standard deviation 12.01. They had attended school from 0 to 36 months, mean 11.52, standard deviation 10.84. Their ages ranged from 61 to 143 months, mean 95.16, standard deviation 23.48. There were 69 boys and 45 girls.

The final battery, which consisted of five tests, was administered to the subjects nine times. The first three administrations yielded pretraining, posttraining, and transfer scores, respectively. The fourth and fifth administrations measured retention, under different conditions and at different times. The last four administrations were repetitions of the first four, given after an interval of seven weeks.

The results demonstrated that the tests were adequate in meeting the criteria established for them. It was also shown that the battery was highly stable. However, as compared with pretraining performance, the learning scores (i.e., posttraining, transfer, and retention) were found to be significantly more stable and significantly more consistent both within and across the tests, suggesting that they measured a more general factor.

Comparisons of the scores on different administrations of the battery demonstrated that the group as a whole learned, transferred the training, and retained the learning to a highly significant extent on all of the tests. It was also shown that the factors of sex and time spent in school did not significantly influence performance on the battery. Age and diagnostic category were found to have a limited effect on the scores of some of the tests.

The variable of intelligence was demonstrated to affect the scores of all of the tests, exerting a highly differential influence both on pretraining as opposed to learning performance, and also on the various kinds of learning performance measured. Transfer and retention scores were found to be the most sensitive measures of differences in ability, suggesting that tests which are administered only once may not be the most discriminating indicators of these differences.

Evidence of construct validity was obtained in the results which showed that learning performance was more stable and more consistent than pretraining performance, in the consistent and stable learning curved obtained by the group on different tests in close accord with the constructs, and in the significantly greater relationship demonstrated between the learning scores and level of ability than between pretraining performance and ability level.

The results also gave partial demonstrations of the predictive validity of the battery. Learning performance, as measured by the tests, was found to be a better predictor than pretraining test performance of intelligence, classroom learning, and social competence, as measured by psychologists' estimated IQs, teachers' ratings, and Vineland SQs.

Within the limitations of the study, the following conclusions are thought to be warranted:

1. The battery was highly stable and consistent.

2. Learning performance was significantly more stable and more consistent than pretraining performance.

3. The factors of sex and time in school did not significantly affect performance on the battery.

4. Age and diagnostic category had a limited effect on the scores of some of the tests.

5. Intelligence exerted a highly significant influence on the scores of all of the tests.

6. Transfer and retention performance were the most sensitive measures of differences in ability.

7. Evidence of construct validity and partial demonstrations of predictive validity for the battery were obtained.

REFERENCES

Arnold, Gwen F. A study of the mental abilities of the cerebral palsied child. Unpublished doctoral dissertation, University of Wisconsin, 1945.

Barnett, C. D., & Cantor, G. N. Discrimination set in defectives. *Amer. J. ment. Defic.*, 1957, **62,** 334–338

Cotton, Carol B. A study of the reactions of spastic children to certain test situations. Unpublished doctoral dissertation, University of Chicago, 1939.

Doll, E. A. *Measurement of social competence.* Philadelphia, Pa.: Educational Test Bureau, 1953.

Edwards, A. L. *Experimental design in psychological research.* New York: Rinehart, 1950.

Guilford, J. P. *Fundamental statistics in psychology and education.* New York: McGraw-Hill, 1956.

Haeussermann, Else. *Developmental potential of preschool children.* New York: Grune & Stratton, 1958.

Harlow, H. F. The formation of learning sets. *Psychol. Rev.*, 1949, **56,** 51–65.

Hilgard, E. R. *Theories of learning.* New York: Appleton-Century-Crofts, 1948

House, Betty, Zeaman, D., Orlando, R., & Fischer, W. *Learning and transfer in mental defectives.* Storrs: Univer. Connecticut, 1957.

Lashley, K. S. The mechanism of vision: XV. Preliminary studies of the rat's capacity for detail vision. *J. gen. Psychol.*, 1938, **18,** 123–193.

Lord, Elizabeth E. *Children handicapped by cerebral palsy.* New York: Commonwealth Fund, 1937.

Penrose, L. S. *Mental defect.* New York: Farrar & Rinehart, 1934.

Pressey, S. L., & Robinson, F. P. *Psychology and the new education.* New York: Harper, 1944.

Reissenweber, Marion. The use of modified block designs in the evaluation and training of the brain-injured. *Psychol. Monogr.*, 1953, **67** (21, Whole No. 371).

Schucman, Helen. A method for measuring educability in severely mentally retarded children: A preliminary study. *Train. Sch. Bull.*, 1957, **54,** 52–54, 58–61; 1958, **55,** 2–4.

Schucman, Helen. *Measurement of the educability of severely mentally retarded children.* Report by New York University to the Office of Education, United States Department of Health, Education, and Welfare, Contract No. SAE-7783, 1959.

Siegel, S. *Nonparametric statistics.* New York: McGraw-Hill, 1956.

Strauss, A. A., & Kephart, N. C. *Psychopathology and education of the brain-injured child.* Vol. II. *Progress in theory and clinic.* New York: Grune & Stratton, 1955.

Strauss, A. A., & Lehtinen, Laura E. *Psychopathology and education of the brain-injured child.* Vol. I. New York: Grune & Stratton, 1950.

Stutsman, Rachel. *Mental measurement of preschool children.* Yonkers: World Book, 1931.

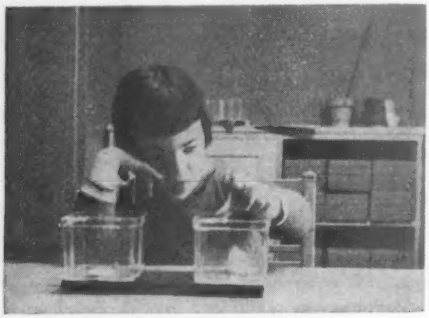
Wallin, J. E. W. *Education of mentally handicapped children.* New York: Harper, 1955.

(Received February 6, 1960)

APPENDIX A

Materials and Procedures for Administering the Tests[A1]

Two-Buttons Test



Materials. Two clear plastic boxes, 3.5 inches high, with sides 3.5 inches along the bottom, and 4 inches along the top. The boxes were attached, 3.5 inches apart, to a black board 10.5 inches long, and 3.5 inches wide.

Eight round white buttons, 1.75 inches in diameter.

Testing Procedures. There were no special pretest procedures for this test, since demonstrations preceded each item. For each item, the subject was required to indicate the box into which each button presented should be placed, in imitation of the examiner's demonstration. If necessary, the examiner picked up the buttons for the subject, and dropped them into the boxes at his indication.

The test consisted of six items, whch are shown on the scoring sheet (Table A1). Each "L" on the scoring blank represents one button to be dropped into the box at the subject's left. Each "R" stands for one button to be put into the box at his right. The first three items required the correct placement of single buttons, one to be dropped into the box at the left, the next into the box at the right, and the third into the box at the left. The fourth item consisted of two buttons, the fifth of three, and the sixth of four. The buttons were dropped into the boxes in alternating order, always beginning at the left.

Boxes were placed about 2 inches from the edge of the table. Buttons were presented approximately an inch from the edge of the table, and 3 inches apart. Before each item was presented, the examiner demonstrated the correct response, dropping the buttons into the boxes from a height of approximately a foot to ensure a sharp click. As the buttons struck the box, the examiner said, "Ping!" in a singing tone, drawing out the final "ng." After dropping each button, the examiner tapped the box into which it had been dropped, and said, "in *here*." Buttons used in the demonstrations were not removed until the subject's response was secured.

After each demonstration, the examiner placed the number of buttons required for the item before the subject, and said, "*You* do it," pointing from the subject to the buttons, and then to the boxes, midway between them. All buttons were removed before the next demonstration. The examiner said, "Ping!" whenever the subject gave a correct response. If the response was wrong, the examiner corrected it, saying, "Ping!" as the buttons were placed correctly.

The first three items were always given. If all three were failed, the test was discontinued, and training was started. Testing was also discontinued if the subject failed the second, third, and fourth items. If he failed the fourth and fifth items, the sixth was omitted.

Training Procedures. Training began with the subject's first failure unless it had occurred on one of the first three items, in which case training began with the first item. Demonstrations were made as before. The examiner then presented the buttons to the subject, saying "in *here*" for each button, pointing to the correct box, and tapping it until the response was made. Again, "Ping!" accompanied all correct or corrected responses.

The subject was trained three times on each item, unless he achieved two consecutive successes. In that event, the third training was omitted. Training was discontinued if there was complete failure on the first three items. Training was also stopped if the second, third, and fourth items were failed. Training on the sixth item was omitted if no successes were achieved on the fourth and fifth.

The Transfer Form. The original items were those used in the transfer situation. However, the demonstrations were omitted. In advance of each presentation, the examiner pointed to the correct box or boxes, saying, "Ping! in *here*," or "Ping! in *here*" and "Ping! in *here*," and so on. As before, correct or corrected responses were accompanied by, "Ping!"

[A1] Photographs were taken at the Shield of David Institute for Retarded Children.

TABLE A1

SCORE SHEET: TWO-BUTTONS TEST

Administration A

(1) L (2) R (3) L (4) L–R (5) L–R–L (6) L–R–L–R Total________

Training

L	R	L	L–R	L–R–L	L–R–L–R
L	R	L	L–R	L–R–L	L–R–L–R
L	R	L	L–R	L–R–L	L–R–L–R

Administration B

(1) L (2) R (3) L (4) L–R (5) L–R–L (6) L–R–L–R Total________

Administration C

(1) L (2) R (3) L (4) L–R (5) L–R–L (6) L–R–L–R Total________

Administration D

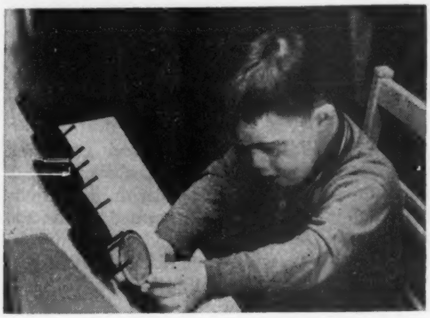
(1) L (2) R (3) L (4) L–R (5) L–R–L (6) L–R–L–R Total________

Administration E

(1) L (2) R (3) L (4) L–R (5) L–R–L (6) L–R–L–R Total________

Comments:

Buzz Test



Materials. A gray board, 23 inches long and 7.5 inches wide, attached at an angle of approximately 45 degrees to a base and a back board, which joined at a 90-degree angle. Four black discs, 1 inch in thickness, and with diameters of 3, 6, 4, and 5 inches, respectively, were attached to the board in that order, with equal distances between them. Thus, beginning at the left, the smallest disc was first, the largest second, the next to the smallest third, and the next to the largest fourth. The sample discs on the board were wired so that, when a disc of corresponding size was placed over one of the samples, a continuous buzzing sound was produced. Two blunt nails projected from the board below each sample, to hold a disc ot comparable size in place.

One set of separate black discs and one set of separate white discs, one disc in each set corresponding in size to each of the sample discs on the board.

Pretest Procedures. The board was placed about 4 inches from the edge of the table. The examiner held the smallest black disc at the center of the board, about 1 inch from the edge of the table, at an angle corresponding to the tilt of the board. After holding the disc in this position for several seconds, the examiner placed it over the corresponding sample on the board, saying, "*This* goes *here*." After allowing the resulting buzz to continue for a few seconds, the examiner returned the disc to the center of the board, holding it as before. Pointing from the subject to the disc, and then to the middle of the board, the examiner said "*You* do it." The examiner placed the disc at the subject's indication, if necessary. One demonstration and one unscored response were allowed for each of the black discs. All errors were corrected.

Testing Procedures. For each item, the examiner held a black disc in the position described above, asking each time, "Where does *this* go?" and pointing from the disc to the board. The subject was required to indicate a choice of one of the samples, the examiner placing the disc for him, if necessary. Errors were corrected before continuing to the next item.

There were six items, each involving matching one black disc to the corresponding sample. The scoring sheet (Table A2) shows the order in which the discs were presented. The number in parenthesis on the scoring sheet refers to the number of the item. The number which follows shows the position on the board of the sample corresponding in size to the disc presented, i.e., "2" represents the largest sample, which was second on the board. All of the items were given, unless the subject failed the first four, in which case the last two were omitted.

Training Procedures. Training always began with the first item. The examiner began by demonstrating the response, saying, "*This* goes *here*," producing the buzz, and then removing the disc and presenting it to the subject, as before. The examiner pointed to the correct sample, and continued to tap it, repeating, "*Here*," until the response was made. If errors still occurred, they were corrected by the examiner.

A special order of the items was used in training on this test, in order to minimize the tendency to perseveration. Instead of the usual procedure of completing training on each item before beginning training on the next, the subject was trained once on all items, in the order in which they were presented in the test, beginning with Item 1. The entire series was repeated three times, unless no errors were made in the course of the first two series.

The Transfer Form. Using the same items, and following the original procedures, the white discs were presented to the subject, to be matched to the black samples of corresponding sizes on the board.

Music Box Test

Materials. Five round music boxes, 2 inches high and 3 inches in diameter. A handle, which could be turned in either direction, projected 2 inches from the center of the box. The shaft of the handle extended vertically from the box for a distance of 1 inch, and was then bent at a right angle, continuing horizontally for another inch, and ending in an upright wooden cylinder, 0.5 inch in diameter and 1 inch long. One box was painted black, one white, one gray, one black and gray, and one gray and white. Only the gray box functioned, playing a repeated series of four tones.

Five rectangular music boxes, 2.5 inches high, 4.5 inches long, and 2.75 inches wide, in other respects like the round boxes.

Pretest Procedures. The round gray box was placed before the subject, about an inch from the edge of the table, with the handle toward him. The examiner turned the handle for some 3 seconds, saying, "music, music, music." The examiner then said to the subject, "*You* make *music*," pointing from the subject to the gray music box. If necessary, the examiner placed the child's hand over or against the handle and turned it for him, repeating, "music, music, music."

TABLE A2

SCORE SHEET: BUZZ TEST

Administration A

(1) 1 (2) 3 (3) 2 (4) 4 (5) 1 (6) 4 Total________

Training

1 3 2 4 1 4 —— 1 3 2 4 1 4 —— 1 3 2 4 1 4

Administration B

(1) 1 (2) 3 (3) 2 (4) 4 (5) 1 (6) 4 Total________

Administration C

(1) 1 (2) 3 (3) 2 (4) 4 (5) 1 (6) 4 Total________

Administration D

(1) 1 (2) 3 (3) 2 (4) 4 (5) 1 (6) 4 Total________

Administration E

(1) 1 (2) 3 (3) 2 (4) 4 (5) 1 (6) 4 Total________

Comments:

The gray box was left on the table, but was pushed aside, some 6 inches to the subject's left. The round black and gray box was presented in center position, its black side toward the gray box. The examiner turned the handle about 3 seconds, shaking his head and saying, "no music, no, no music." The subject was then asked to turn the handle, as before, the examiner repeating, "no music" as he responded.

The black and gray box was pushed to the subject's right, and the round gray and white box was presented, with the white side toward the gray box. The demonstration and response were repeated. The box was then pushed to the subject's left, and the gray box was presented again. The examiner played it, again for about 3 seconds, saying *"Here's* the music," and asking the child to play it, as before. The boxes were removed before the test began.

Testing Procedures. Boxes were presented about 4 inches apart, the handles turned toward the subject. At each presentation, the examiner said, "*You* make music," pointing from the child to the music boxes. The subject was required to indicate his choice, the examiner turning the handle of the box he selected, if necessary. If the subject's response was wrong, the examiner pointed to the gray box, turned the handle, and said, "*Here's* the music."

The shifts in the positions of the boxes were made with the subject watching. However, to prevent uninterrupted visual fixation on the gray box, the subject's attention was briefly diverted during each shift. To keep the interruption as standard as possible, the examiner said, "Look at me, (subject's name)," and then redirected the child's attention to the table. This required 2 or 3 seconds.

The scoring sheet (Table A3) shows the

TABLE A3

Score Sheet: Music Box Test

Administration A

(1) G–B (2) W–G (3) G–W–B (4) B–W–G–W/G (5) W/G–G–W–B–B/G
(6) W–B–B/G–G–W/G Total______

Training

G–B	W–G	G–W–B	B–W–G–W/G	W/G–G–W–B–B/G	W–B–B/G–G–W/G
G–B	W–G	G–W–B	B–W–G–W/G	W/G–G–W–B–B/G	W–B–B/G–G–W/G
G–B	W–G	G–W–B	B–W–G–W/G	W/G–G–W–B–B/G	W–B–B/G–G–W/G

Administration B

(1) G–B (2) W–G (3) G–W–B (4) B–W–G–W/G (5) W/G–G–W–B–B/G
(6) W–B–B/G–G–W/G Total______

Administration C

(1) G–B (2) W–G (3) G–W–B (4) B–W–G–W/G (5) W/G–G–W–B–B/G
(6) W–B–B/G–G–W/G Total______

Administration D

(1) G–B (2) W–G (3) G–W–B (4) B–W–G–W/G (5) W/G–G–W–B–B/G
(6) W–B–B/G–G–W/G Total______

Administration E

(1) G–B (2) W–G (3) G–W–B (4) B–W–G–W/G (5) W/G–G–W–B–B/G
(6) W–B–B/G–G–W/G Total______

Comments:

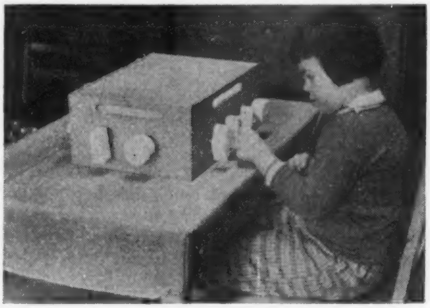
number of boxes used for each of the six items, and the order in which they were presented. The letter "G" on the sheet represents the gray box, "B" the black, "W" the white, "B/G" the black and gray, and "W/G" the white and gray. Two boxes were used for each of the first two items. Three were used for the third, and four for the fourth. Five boxes were used for the fifth and sixth items. The two-toned boxes were not introduced into the test until the fourth item. One of them was used in the fourth item, and both in the fifth and sixth. In the fifth item only one of the two-toned boxes was placed next to the gray box. In the sixth item, the gray box was presented with a two-toned box on each side. The two-toned boxes were always presented with the gray side away from the gray box. The first three items were always given. The test was discontinued if these three items, or the second, third, and fourth items were failed. The sixth item was not given if the subject failed the fourth and fifth.

Training Procedures. If the subject's first failure occurred on one of the first two items, training began with the first. Otherwise, it began at his first failure. The examiner presented the item, turned the handle of the gray box, and said, "music, music, music." Then, pointing from the subject to the gray box, the examiner said, "*You* make *music,*" tapping the gray box, and repeating, "*Here's* the music," until the response was secured. If necessary, the examiner placed the handle of the gray box into or against the subject's hand, turning the handle for him, and saying, "Music, *here's* the music." No interruptions were made in the shifts of materials during training.

The subject was trained three times on each item, or through two consecutive successes. If all test items were failed originally, training was always given on the first three items. It was discontinued if these were failed completely, or if complete failure occurred in training on the second, third, and fourth items. Training was omitted for the sixth item, if no successes were achieved on the fourth and fifth.

The Transfer Form. The rectangular music boxes were used in place of the round ones. The original testing procedures were followed, and the original test items were used. The round set was pushed to the back of the testing table, but was not removed, in an attempt to emphasize the fact that different boxes were being presented.

Light Test



Materials. Two gray boxes, each 14.5 inches long, 11 inches wide, and 6.5 inches high. White knobs were used on one of the boxes, and corresponding black knobs on the other. Above the knobs, on each side of the box, was a plastic covered "window," 6 inches long and 1 inch wide. Behind the windows were a series of flashlight bulbs, which lit when the rectangular knob was pressed. A single rectangle appeared on each side. A rectangle and a sphere were used on Side 1. Side 2 contained a diamond and a rectangle. Side 3 included a rectangle, a diamond, and a triangle. A diamond, a rectangle, and a triangle, and a sphere were used on Side 4. The spheres were 2 inches in diameter. The triangles were 2.5 inches on each side. The diamonds were 2.75 inches long, and 1.5 inches wide. The rectangles were 2.5 inches high and 1 inch wide.

Pretest Procedures. The box with white knobs was placed about 2 inches from the edge of the table, Side 1 facing the subject. The examiner pressed the rectangular knob, pointed from the knob to the lighted window, and said, "*Light,* see the *light.*" Then the examiner pressed the sphere, pointed from the knob to the unlit window, shook his head, and said, "No light, see, no light." Next, the examiner pressed the rectangle again, pointing as before, and saying, "*Here's* the light." Turning to the subject, the examiner said, "*You* make *light,*" pointing to the subject, and then to the box. If necessary, the examiner pressed the knob at the subject's indication, and corrected the response in case of error. As the rectangle was pressed, the examiner pointed again to the window, repeating, "See the *light.*" The examiner then turned the box so that Side 2 faced the subject, and the entire demonstration was repeated.

Testing Procedures. The test contained eight items, each consisting of the presentation of one side of the box. The order in which the sides were presented is shown on the scoring sheet (Table A4). As each side was presented, the examiner said, "*You* make *light,*" pointing from the subject to the box. If necessary, the examiner pressed the knob which the subject indicated as his choice, and corrected any errors made. All items were given unless the subject failed the first four.

Training Procedures. Training began with Side 1. The examiner pressed the rectangular knob, saying, "*Here's* the *light,*" pointing from the rectangle to the lighted window. The subject was then asked "to make light," the examiner pointing to the rectangular knob and tapping it until the subject's response was made. The examiner repeated "*Here's* the light" as the rectangular knob was pressed, and corrected errors if necessary.

The subject was trained three times, or through two consecutive successes, on each side of the box. Training was given on each of the four sides, unless the subject achieved no successes in training on the first three sides.

The Transfer Form. The original procedures were followed, using the box with black

TABLE A4

SCORE SHEET: LIGHT TEST

Administration A

(1) 1 (2) 2 (3) 3 (4) 4 (5) 1 (6) 2 (7) 3 (8) 4 Total________

Training

Side 1	Side 2	Side 3	Side 4
Side 1	Side 2	Side 3	Side 4
Side 1	Side 2	Side 3	Side 4

Administration B

(1) 1 (2) 2 (3) 3 (4) 4 (5) 1 (6) 2 (7) 3 (8) 4 Total________

Administration C

(1) 1 (2) 2 (3) 3 (4) 4 (5) 1 (6) 2 (7) 3 (8) 4 Total________

Administration D

(1) 1 (2) 2 (3) 3 (4) 4 (5) 1 (6) 2 (7) 3 (8) 4 Total________

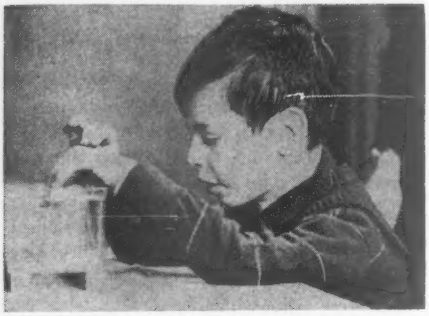
Administration E

(1) 1 (2) 2 (3) 3 (4) 4 (5) 1 (6) 2 (7) 3 (8) 4 Total________

Comments:

knobs. The first box was pushed to the back of the testing table, but was not removed from sight.

Three-Buttons Test



Materials. Three clear plastic boxes, 3.5 inches high, 3.5 inches along the bottom, and 4 inches along the top. The boxes were placed 3 inches apart, and attached on parallel sides to two wooden runners, 17 inches long, 1 inch wide, and 1 inch high.

Nine round buttons, 1.5 inches in diameter, 3 black, 3 white, and 3 gray.

Pretest Procedures. The boxes were placed about 2 inches from the edge of the table. The examiner dropped a sample white button into the box at the subject's left, saying, "Ping!" as the button struck the bottom of the box. Leaving the sample in place, the examiner placed another white button before the subject, about 1 inch from the edge of the table and approximately at the center point of the middle box, and said, "*You* do it," pointing from the subject to the button, and then to the box containing the sample. If necessary, the examiner placed the button at the subject's direction, and corrected or correctly completed the response. These procedures were repeated with black and gray buttons, using the middle and the right hand boxes, respectively. The examiner added "Ping!" as each button was correctly placed. The sample buttons remained in the boxes throughout testing and training.

Testing Procedures. Buttons were presented about an inch from the edge of the table. Single buttons were presented approximately at the center of the boxes. For the two-button items, one button was placed at the mid-point of the space between the left and center boxes, and the other at the corresponding point between the right and center boxes. For the final item, the middle button was placed at the center of the boxes, and the others at the mid-points of the spaces between the center box and those to the left and right, respectively.

The buttons used for each item were removed after its completion. Buttons used in each of the six items are shown on the scoring sheet (Table A5). "W" represents a white button, "B" a black one, and "G" a gray one. The first three items consisted of placing a single white, black, and gray button into the box containing the corresponding sample, the examiner dropping the buttons into the boxes at the subject's direction, if necessary. The fourth item consisted of a white and a gray button, and the fifth of a gray button and a white one, presented in that order. A black, a gray, and a white button were used for the sixth item. A correct response for the latter

TABLE A5

SCORE SHEET: THREE-BUTTONS TEST

Administration A

(1) W (2) B (3) G (4) W–G (5) G–W (6) B–G–W Total________

Training

W	B	G	W–G	G–W	B–G–W
W	B	G	W–G	G–W	B–G–W
W	B	G	W–G	G–W	B–G–W

Administration B

(1) W (2) B (3) G (4) W–G (5) G–W (6) B–G–W Total________

Administration C

(1) G (2) W (3) B (4) G–B (5) B–G (6) W–B–G Total________

Administration D

(1) W (2) B (3) G (4) W–G (5) G–W (6) B–G–W Total________

Administration E

(1) W (2) B (3) G (4) W–G (5) G–W (6) B–G–W Total________

Comments:

items required the correct placement of all of the buttons presented.

As each item was presented, the examiner asked, "Where does *this* go?" or, "Where do *these* go?" pointing from the buttons to the boxes. The examiner accompanied each correct placement with a "Ping!" Errors were corrected, or correctly completed by the examiner. The first three items were always presented. Testing was discontinued if the subject failed the first three items, or the second, third, and fourth. The last item was omitted if the fourth and fifth were failed.

Training Procedures. Training was given three times on each item, or through two consecutive successes. Training always began with the first item. However, if the subject had passed the earlier items on the test, he was trained only once on those items, regular training beginning at the point of his first failure.

Before each item, the examiner demonstrated the correct placement for each button, presented the item, and pointed to the correct box until the subject had responded, saying, "in *here*," while tapping the edge of the box. All errors were corrected, and a "Ping!" accompanied a correct response made by the subject, or by the examiner at the subject's indication. Training was discontinued if there was complete failure on the first three items, or on the second, third, and fourth. Training was omitted on the sixth item if the subject failed the fourth and fifth.

The Transfer Form. The samples for the original form of the test were removed, and those used for the transfer situation were placed into the boxes by the examiner, following the procedures used in placing the original samples. The positions of the samples, and the buttons used for each item, are shown in Table A5 (see Administration C). The original testing procedures were followed.

APPENDIX B

The Experimental Test Battery

This section describes the preliminary battery used in the pilot studies. The various revisions of the tests which were ultimately included in the final battery are described first. The rejected tests are described thereafter.

Revisions of Tests Used in the Final Battery

The *Two-Buttons Test* underwent eight revisions in the pilot studies. The test originally contained five items. The first two consisted of single buttons, one to be placed in the box to the subject's left, and the other in the box to his right. Two, three, and four buttons were then presented, to be dropped into alternate boxes. The resulting scores were too low. The earlier pilot studies therefore concentrated on adding various easy items.

The addition of two single-button items was attempted first. This, however, produced a marked increase in perseveration. When the new items required reversing the order of placement used for the original items, the subject found the shift almost impossible to make, and training merely tended to increase this difficulty rather than to overcome it. On the other hand, when the added items duplicated the original order, the subjects continued using that order regardless of the number of buttons presented. Perseveration decreased to a satisfactory extent when the number of single-button items was reduced to three, and only an alternating order was required for all placements.

Results of various changes in the multiple-button items were fairly consistent. Items containing odd numbers of buttons were found to be more difficult than even-numbered items, to the extent that training for a three-button item tended to be less successful than training for an eight-button item. However, once a subject had grasped the concept of placing buttons in alternating order, increasing the buttons in even numbers presented little additional difficulty. Items consisting of six and eight buttons were therefore dropped as superfluous.

Of the various odd-numbered items attempted, only one three-button item was retained, since inclusion of two items using odd numbers of buttons made the test too difficult. The three-button item was tried out in various positions, and was found to be most useful in next-to-last position.

Various transfer forms for this test were also studied. In some, the examiner used the original white buttons for the demonstrations, presenting black and/or gray buttons to the subject. These were found to be too easy, since higher scores than on the retest administration were obtained. On the other hand, when black and white buttons were both presented to the subject, the resulting scores were too low. Omission of pretest demonstrations and instructions also made the form too difficult, the subjects apparently failing to understand what was required of them. However, omitting the demonstrations, but pointing to the correct box or boxes in advance of the presentations, seemed to provide sufficient cues.

Of the various sounds tried out as the verbal accompaniment of the demonstrations, "Ping!" was found to be the most interesting to the children, especially when pitched high and given a drawn-out "ng" at the end. Virtually all of the subjects smiled or laughed at this sound. Because of its evident appeal, the sound was reserved as an accompaniment for correct responses only, to increase motivation and reinforce learning.

The *Buzz Test* was revised six times in the preliminary experimentation. As it was originally constructed, the sample discs were arranged on the board in order of size. A round gray peg projected at right angles from the center of each sample, and extending 2 inches beyond it. Each of the discs presented to the subject had a round hole, 1.5 inches in diameter, at the center. The discs slipped over the pegs projecting from the samples, the buzz being produced by pressing a disc against its corresponding sample on the board. The majority of the subjects were unable to make the necessary size discriminations for this version of the test. It appeared to be especially difficult for the brain-injured children.

The pegs were removed in the next pilot study, and discs without holes were used. To produce the buzz, the presented disc was laid over the corresponding sample, supported by blunt nails driven into the board just below the samples. This change made the test easier, and did not appear to penalize the brain-injured subjects. However, perseveration, especially after training, was so great that some of the subjects did not shift from the first response for the remainder of the test. It is interesting to note that perseveration occurred to a significant extent only on this test and the preceding one. It was virtually absent on all other tests throughout their revisions.

In an attempt to overcome perseveration, the differences between the sizes of the discs were increasd. This was unsuccessful. In the next attempt, the arrangement of the samples was changed, to avoid a regular increase in size. The smallest disc was placed first and followed by the largest. The next to the smallest came next, with the next to the largest, last. Perseveration was almost entirely overcome. However, the test was now too easy.

To increase the difficulty, presenting each disc twice was attempted. Perseveration increased sharply. Presenting the discs in the order in which they were used in the test, and then repeating the series again in that order, overcame the tendency to perseveration, but made the test too long to hold interest. Two items were therefore dropped. One complete series was given, plus two additional items consisting of one placement at the extreme left and one at the extreme right. This did not give rise to perseveration, and shortened the test sufficiently.

The *Music Box Test* was changed six times. In its original form, it consisted of two gray boxes and two white boxes. The one functioning box was black. Since this form was found to be too easy for purposes of discrimination, two additional boxes were used, one gray and one white. However, this increase in the length of the test tended to make the subjects restless. The number of boxes was reduced again to five, and one of the gray boxes was darkened considerably. No marked increase in difficulty resulted. However, when the boxes were repainted so that the darker gray box played, the subjects found it too difficult to distinguish between the two shades of gray.

A two-toned box, half white and half black, was introduced in the next attempt, the single gray box playing. Again the test was found to be too easy. Two two-toned boxes, one black and gray and one gray and white were tried next, presented with their gray sides facing the gray box. This tended to confuse the subjects, especially the brain-injured children. In the final form, the two-toned boxes were always presented with their gray sides away from the gray box, which considerably lessened the confusion, and also made the test more suitable for the brain-injured subjects. Since the original rectangular boxes were satisfactory for the transfer form, no changes were made in their size and shape. However, they were painted to correspond with the round boxes in each revision.

The *Light Test* underwent seven revisions. Only three shapes were used in its original construction: spheres, triangles, and rectangles, the latter being the correct response. Three knobs were the maximum used on any one side of the box. Although this form was too easy for the group as a whole, the first two sides were not changed, the simpler items being necessary for discrimination among the more retarded subjects. The revisions therefore concentrated on increasing the difficulty of the remaining items.

A larger box was constructed. The fourth side now containing four knobs, including a second sphere. Since this did not increase the difficulty sufficiently, all spheres were removed from the fourth side, and a diamond was introduced. This was still too easy. One sphere was therefore added, the rectangle being inserted between the sphere and the diamond. Since this also was unsuccessful, the rectangle was placed between the diamond and the triangle, a change which greatly increased the difficulty of the item. The original sphere on the third side of the box was also replaced by a diamond, which increased the difficulty of this item as well.

The Light Test was so interesting to the subjects that the length of the test could be increased without much risk of loss of interest. Two methods were tried for adding to the number of the items. In the first method, each side of the box was presented twice, before the next side was turned to the subject. This produced highly uneven scores, the second item on each side being considerably easier than the first. In the final revision, the sides were presented in the order of difficulty, the entire procedure being repeated twice.

The *Three-Button Test* was revised four times in the course of the pilot studies. In its original form, in which only single buttons were used, it was found to be too easy for purposes of discrimination. Introducing two-button items did not increase the difficulty sufficiently. On the other hand, proceeding directly from single buttons to three-button items made the test too difficult. In the final form, three single-button presentations were used, followed by two two-button items, with one three-button item as the final presentation.

Rejected Tests

The *Sounds Test,* a test of shape discrimination, consisted of a gray board like that used in the Buzz Test, but 2 inches longer. Projecting from the board were four knobs: an octagon, a triangle, a gear, and a star. Pressing the octagon caused chimes to sound, the triangle rang a bell, the gear produced a buzz. Pressing the star caused no sound. The subject was shown, in pretest demonstrations, the results of pressing the various knobs. He was then asked to reproduce the sound, being instructed "to make ting-a-ling," "to make buzz," and so on. For the star, he was asked "to make sh-sh," the examiner putting a finger to his lips to indicate silence. Various patterns of sounds were used during the pilot studies.

A number of special difficulties arose. One problem was the unwillingness of the subjects to shift from one knob to another. This was apparently not due to a perseverative tendency, since the subjects seemed to he able to shift when they elected to do so. However, they liked the sounds, and, having produced one, they tended to repeat it with evident enjoyment, regardless of instructions.

In an effort to force the shift, cut-off buttons were installed, so that the examiner could shut off the sound after the response was secured. This tended both to confuse and to anger the subjects, who often pressed harder and harder on the knob, apparently to "make it work." In the next attempt, the examiner demonstrated the next response immediately after completion of the previous one, thus directing the child's attention to the new sound. This was successful in facilitating shifts, but reduced the test to a simple imitation problem.

All attempts to secure responses based on sound patterns were apparently beyond the abilities of the subjects, even when only two sounds were used at a time. The test was finally abandoned as too difficult for purposes of discrimination for this population. It was tried out experimentally with a group of moderately retarded subjects, for whom it appeared to be highly suitable. The subjects enjoyed the test, and were able to handle simple sound patterns. A transfer form based on reversals of the original patterns also appeared to be feasible. Since the test can easily be used for presenting patterns ranging from the very simple to the highly complex, it might serve in an upward extension of this method of testing.

The *Rattle Test* was a test of imitation, in which subject and examiner used comparable sets of material. The subject was required to select the correct material, and use it in imitation of the examiner's demonstration. The materials included a rattle, a small pillow, a hairbrush, and a cup. After each demonstration, the subject was asked to "make rattle, rattle, rattle," "to go to sleep," and so on. A correct response required that the subject indicate the material to be used, and the part of the body with which its use was associated, i.e., the head for the pillow, the mouth for the cup, and so on.

This test was so appealing to the subjects that it was retained in various forms until the last pilot study. Unfortunately, the revisions did not succeed in making it sufficiently useful for purposes of discrimination. When the materials were presented singly or in groups of twos or threes, the test was too easy. When more complicated patterns were attempted, the test became too difficult. Similar difficulties were found in the various transfer forms. Reversals of the order of the materials was too easy, while the use of the verbal instructions without the demonstrations was too difficult.

Though the seven revisions failed to result in an acceptable form of the test, it is still thought to be potentially useful because of its exceptional appeal to the children. It is also thought to have possibilities for other groups of retarded subjects. Limited experimentation suggested that presentation of the materials singly and in small groups would be suitable for younger severe retardates, while more complicated patterns seemed to be appropriate for the moderately retarded. The test was highly successful in holding the interest of both groups.

The *Bang Test* was also a test of imitation. The examiner struck the table an increasing number of times, saying, "Bang!" at each strike. The subject was then asked "to make bang!" or "to make bang! bang!" and so on. The same procedures were repeated for the transfer form, using the vertical surface of a wall instead of the horizontal tabletop. Various items in different orders were tried, the number of strikes ranging from one to seven.

The test presented several problems. A major difficulty was a marked tendency on the part of the subject to respond with random banging, a tendency which training did not often overcome. However, the most serious problem was that a number of the subjects disliked the test. Although they appeared to enjoy the demonstrations, they did not want to imitate them. In the case of subjects with whom the training was successful, another type of difficulty arose: some continued to strike or bang the material of subsequent tests, disrupting the testing situation.

The *Drum Test* and the *Tambourine Test* were attempts to retain the method of the preceding test. The items in the Bang Test were repeated, first by beating against a drum with the hand, and later by striking a tambourine. In both cases, the subjects did enjoy the test. However, their tendency to respond randomly was not lessened, and attempts at training tended to increase it still further.

The *Bell Test* was the most successful of the attempts to utilize the principles of the Bang Test. A domed white bell was used, in which a clapper hit the metal side of the bell when the projecting top was struck. The bell was very appealing to the subjects, though several revisions were necessary in its construction before it was suitable. A flat white disc raised well above the bell's dome was used in place of the original top, which was too small to attract the subjects' attention. More successful was a black disc, set high above the dome.

The use of the bell decreased the tendency to respond randomly, although it did not overcome it entirely. However, now that a greater number of correct responses were secured, another problem came into evidence. Subjects who had learned rudimentary counting performed better than subjects who had not. The test was therefore abandoned, since successful performance so evidently depended on differential past experience.

The *Egg Test* was a test of size discrimination. A lure was placed in the largest of a series of wooden eggs of various sizes. This egg was then presented to the subject in a different position for each item, the eggs being used in increasing numbers. Each egg was presented resting on its base, tapered end upward.

Since the first form of the test was found to be too easy, decreasing the differences between the sizes of the eggs was attempted. This change made the test too confusing, especially for the brain-injured subjects. Several attempts were also made to present the egg in a horizontal position. A gray board with oval indentations was used to prevent the eggs from rolling out of position. These attempts were not successful. When the eggs were presented so that the line from base to apex was parallel to the edge of the table, the test was too easy. On the other hand, the test became too difficult when either the tapered ends of the eggs or their bases were presented toward the subject. Several transfer forms were attempted, using black eggs of the same sizes as the white set, and various other shapes of corresponding relative sizes.

The test was finally abandoned as it became increasingly evident that the Buzz Test was a more suitable size discrimination test for this population, and was also easier to administer. However, the Egg Test was attempted in its first form with a group of younger severely retarded subjects, for whom the Buzz Test was too difficult. Results suggested that the Egg Test may be suitable for the younger subjects.

mina-
of a
This
in a
s be-
egg
l end

nd to
tween
This
cially
empts
hori-
l in-
from
were
ented
rallel
easy.
fficult
gs or
bject.
using
e set,
rela-

ecame
was a
this
ister.
in its
erely
t was
Egg
jects.